Politecnico di Torino

DISMA ECCELLENZA 2018·2022

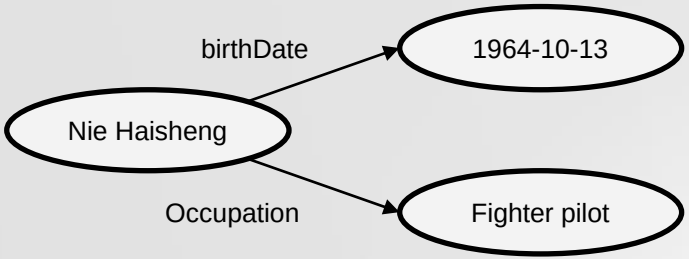
Hybrid Neural Knowledge Graph-to-Text and Text-to-Text Generation

Master's degree in Mathematical Engineering

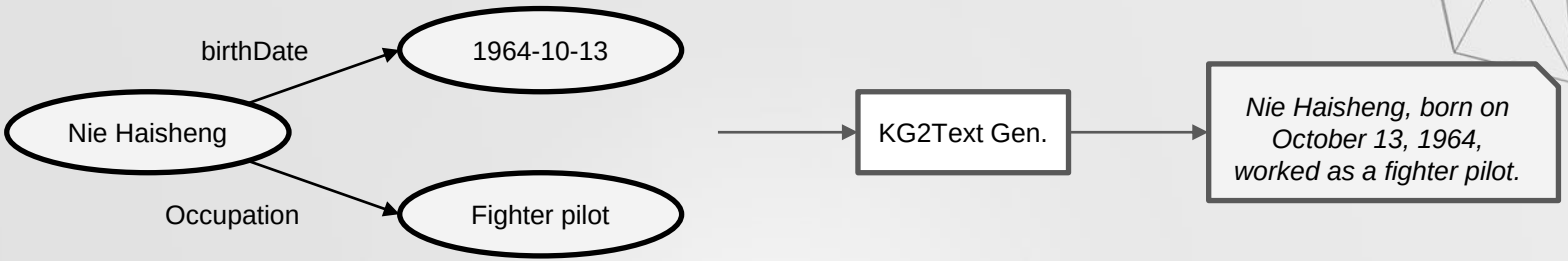
Candidate: Marco Saponara

Academic supervisor: Tatiana Tommasi
External supervisor: Leo Wanner

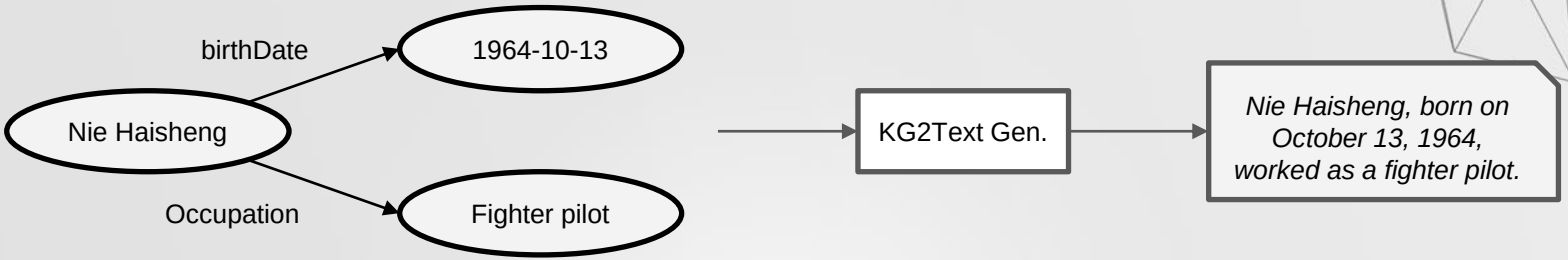
Our task (1/2)


birthDate
1964-10-13
Nie Haisheng
Occupation
Fighter pilot

Our task (1/2)


birthDate
1964-10-13
Nie Haisheng
Occupation
Fighter pilot
KG2Text Gen.
Nie Haisheng, born on October 13, 1964, worked as a fighter pilot.

Our task (1/2)



Nie Haisheng

From Wikipedia, the free encyclopedia

In this Chinese name, the family name is Nie.

Nie Haisheng (born 13 October 1964[*citation needed*]) is a major general of the Chinese People's Liberation Army Strategic Support Force (PLASSF) in active service as an taikonaut and the third commander (unit chief) of the PLA Astronaut Corps (PLAAC). He was a PLA Air Force fighter pilot and director of navigation.

Nie flew on Shenzhou 6 and served as commander on both the Shenzhou 10 and Shenzhou 12 missions, the latter of which became the first crew to visit the Tiangong space station. With a combined 111 days in space, in 2021 he set a new record for longest stay in space by a Chinese astronaut, and is one of only two Chinese astronauts to have flown three times.[2]

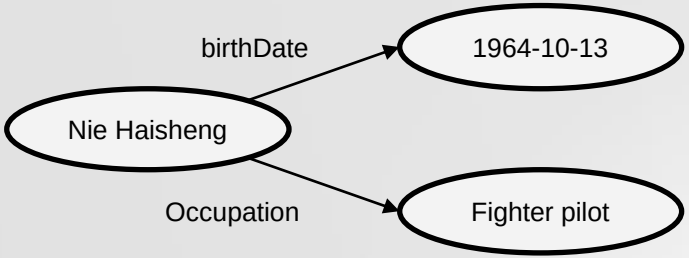
Contents [hide]

1 Air Force career
2 Astronaut Corps career
3 Personal
4 See also
5 References
6 External links

Air Force career [edit]

Nie was born on 08 September 1964[3][4] in Yangdang Town of Zaoyang County, Xiangyang City, Hubei Province. After graduating from high school he joined the People's Liberation Army Air Force in June 1983, and became a fighter pilot. He trained at the PLAAF's No. 7 Flying School and graduated in 1987.[5]

Our task (1/2)



Nie Haisheng

From Wikipedia, the free encyclopedia

In this Chinese name, the family name is Nie.

Nie Haisheng (born 13 October 1964[*citation needed*]) is a major general of the Chinese People's Liberation Army Strategic Support Force (PLASSF) in active service as an taikonaut and the third commander (unit chief) of the PLA Astronaut Corps (PLAAC). He was a PLA Air Force fighter pilot and director of navigation.

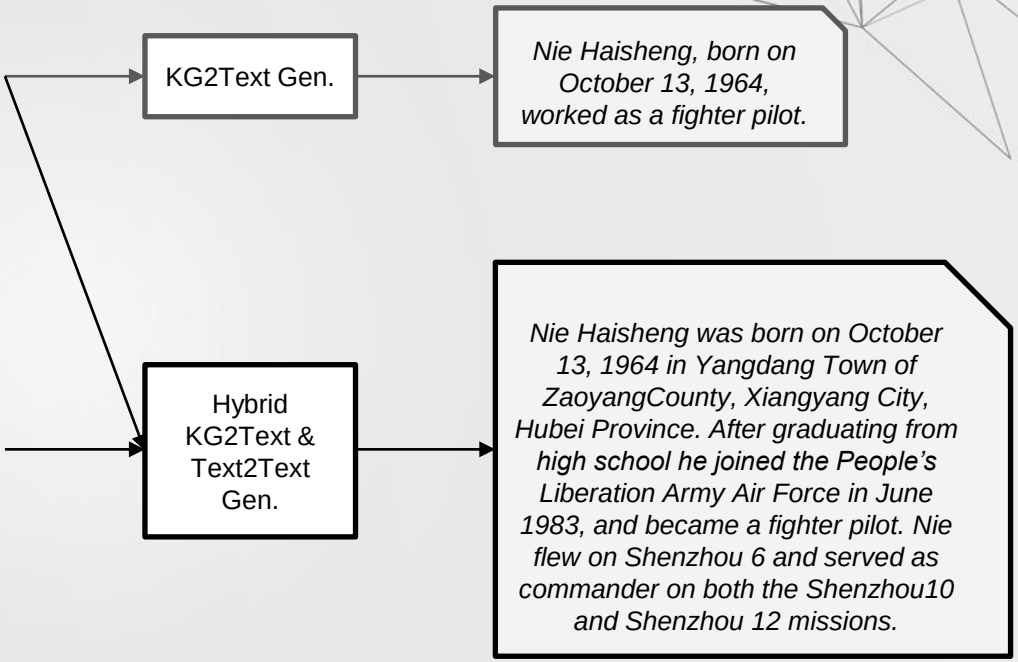
Nie flew on Shenzhou 6 and served as commander on both the Shenzhou 10 and Shenzhou 12 missions, the latter of which became the first crew to visit the Tiangong space station. With a combined 111 days in space, in 2021 he set a new record for longest stay in space by a Chinese astronaut, and is one of only two Chinese astronauts to have flown three times.[2]

Contents [hide]

1 Air Force career
2 Astronaut Corps career
3 Personal
4 See also
5 References
6 External links

Air Force career [edit]

Nie was born on 08 September 1964[3][4] in Yangdang Town of Zaoyang County, Xiangyang City, Hubei Province. After graduating from high school he joined the People's Liberation Army Air Force in June 1983, and became a fighter pilot. He trained at the PLAAF's No. 7 Flying School and graduated in 1987.[5]



Our task (2/2)

MOTIVATION

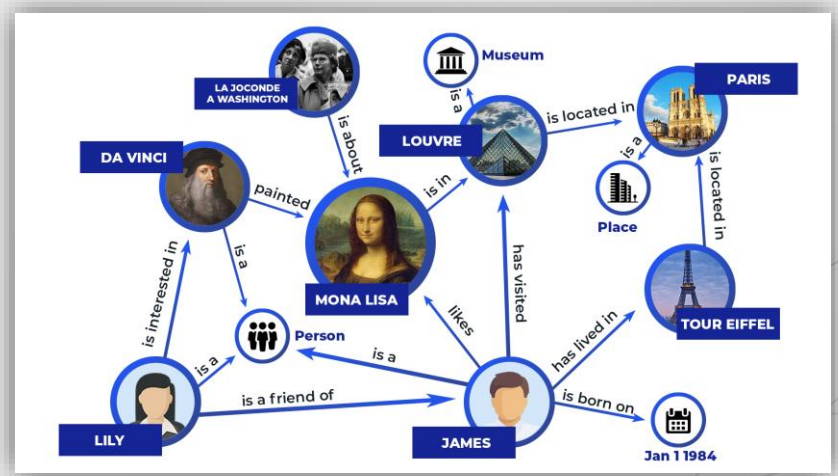
- Knowledge graphs cover a small amount of facts
- Additional information is available in textual format, but its semantic representation’s extraction is expensive

CHALLENGES

- Combination of input at different levels of linguistic abstraction
- Lack of training data and documentation
- Assessing the quality of a generated text in a referenceless framework

Background – Resource Description Framework (RDF)

- Standard data model for data interchange on the Web
- Knowledge representation language
- It facilitates data merging from different data sources
- Statements about resources are represented through **subject-predicate-object** expressions, i.e., the **semantic triples**
- A collection of RDF statements forms a **labeled, directed graph**



Background – Language modeling

- Language modeling = next-word prediction problem

Background – Language modeling

- Language modeling = next-word prediction problem
- Each sentence is treated as a sequence of words:

$$x = (x_1, x_2, \dots, x_n)$$

- The probability of x is decomposed with the **chain rule**:

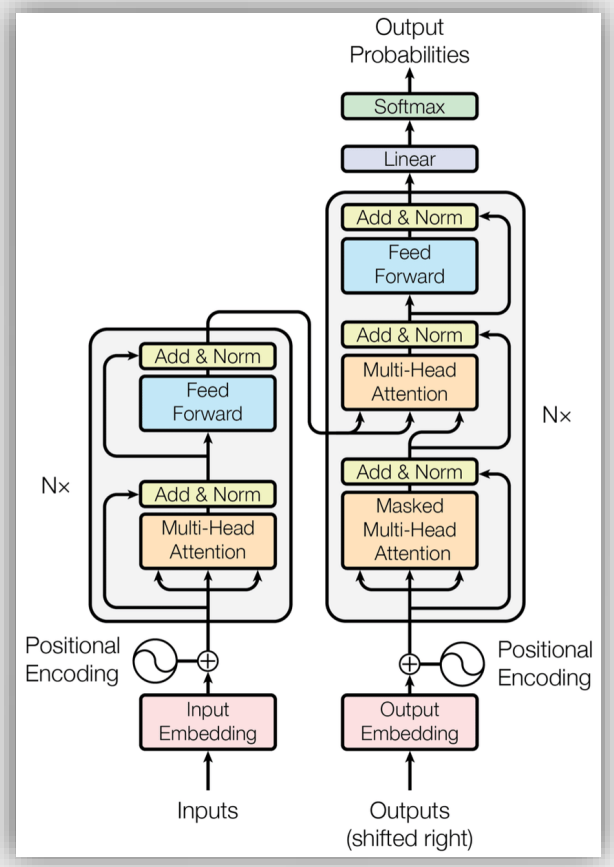
$$p(x) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{<i})$$

- A language model with parameters θ is trained over a textual corpus D by maximizing the log-likelihood:

$$L(\theta) = \sum_{i=1}^{|D|} \log p_\theta(x^i)$$

Background - Transformer

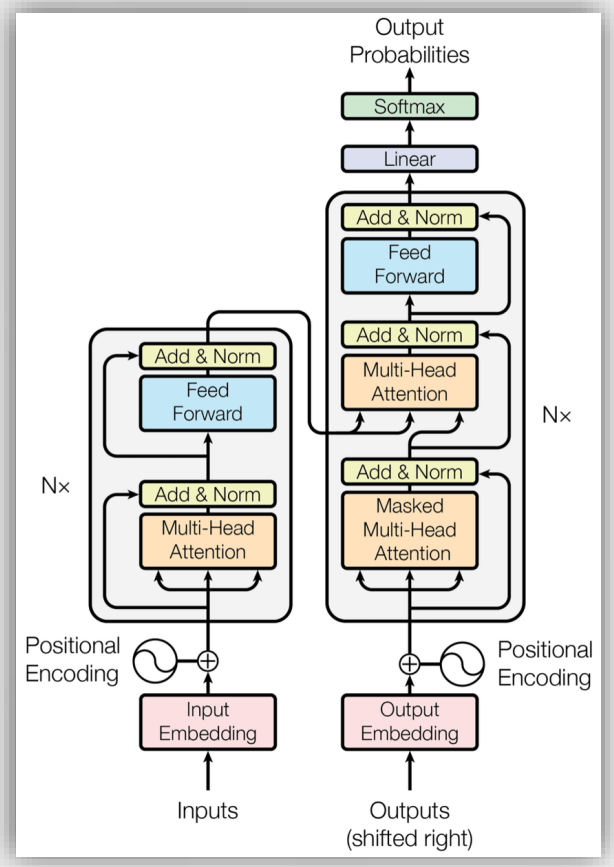
- Sequence-to-Sequence (Seq2Seq) neural architecture
- Encoder-Decoder model
- It relies on the concept of **self-attention**
- Self-attention allows to capture the relationships between each part of a sequence by weighing differently each part of the input data according to their relative importance



Background - Transformer

- Sequence-to-Sequence (Seq2Seq) neural architecture
- Encoder-Decoder model
- It relies on the concept of **self-attention**
- Self-attention allows to capture the relationships between each part of a sequence by weighing differently each part of the input data according to their relative importance

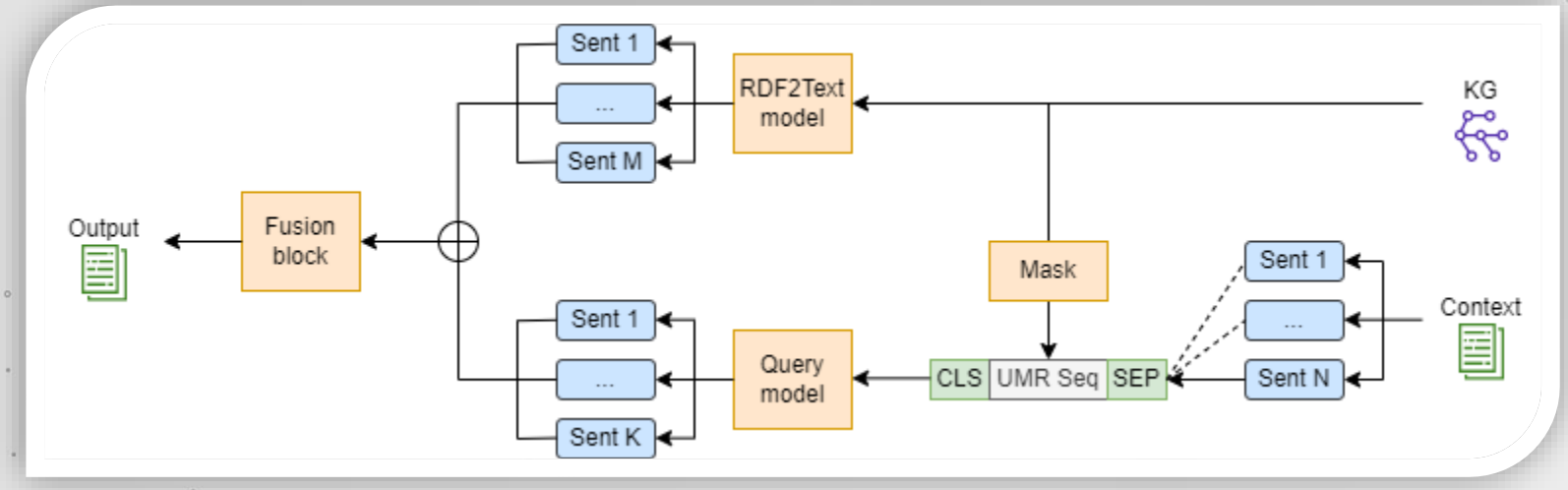
- Transformer-based models employed in our work: **T5, BERT**



Our system

Based on three steps:

1. pure RDF-to-Text generation
2. content selection from the context
3. combination of the intermediate outputs from the previous steps



Our system – RDF-to-text generation

Model:

- Off-the-shelf model from the **WebNLG-2020** Challenge
- Approach based on **transfer learning**
- Pretrained **T5** model fine-tuned on the WebNLG English Dataset

Our system – RDF-to-text generation

Model:

- Off-the-shelf model from the **WebNLG-2020** Challenge
- Approach based on **transfer learning**
- Pretrained **T5** model fine-tuned on the WebNLG English Dataset

Dataset:

- Data/text pairs where the data is a set of triples from **DBpedia** and the text is a verbalization of these triples
- 16 categories: *Airport*, *Astronaut*, *Building*, *City*, *ComicsCharacter*, *Food*, *Monument*, *SportsTeam*, *University*, *WrittenWork*, *Athlete*, *Artist*, *CelestialBody*, *MeanOfTransportation*, *Politician*, and *Company*

Our system – Content selection from the context

Model:

- **BERT**-based regression model trained on custom dataset
- It measures the relevance score of a given sentence from the context w.r.t. the associated RDF triples
- At inference time, it computes the relevance score for each sentence in the context, ranks them and selects the top-k sentences

Our system – Content selection from the context

Model:

- **BERT**-based regression model trained on custom dataset
- It measures the relevance score of a given sentence from the context w.r.t. the associated RDF triples
- At inference time, it computes the relevance score for each sentence in the context, ranks them and selects the top-k sentences

Dataset:

- KG/context pairs where the KG derives from the **WebNLG** dataset and the context is the collection of **Wikipedia** articles associated with the subjects in the triples
- After preprocessing, the input is the concatenation of a processed version of the triples and a sentence from the context, the target is the relevance score
- The relevance score is derived from the **ROUGE** score (a metric based on n-gram overlap) between the sentence and the tripleset

Our system – Fusion block

Model:

- It aims at combining the KG's verbalization and the selected sentences into a fluent and cohesive text (→ sentences' reordering, repetitions' removal)
- Approach based on **transfer learning**
- Pretrained **T5** model fine-tuned on custom Dataset

Our system – Fusion block

Model:

- It aims at combining the KG's verbalization and the selected sentences into a fluent and cohesive text (→ sentences' reordering, repetitions' removal)
- Approach based on **transfer learning**
- Pretrained **T5** model fine-tuned on custom Dataset

Dataset:

- Text pairs derived from the target texts in the WebNLG Dataset
- The input text is a noisy version of the target

Results – Examples (1/3)

BASELINE

The Acharya Institute of Technology can be found in India in the state of Karnataka, is affiliated with Visvesvaraya Technological University, and the campus address is: Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore - 560090.

OUR OUTPUT

The Acharya Institute of Technology is located in the state of Karnataka in India. It is affiliated with the Visvesvaraya Technological University and has a Campus in the academic discipline of Mechanical engineering and is affiliated with the All India Council for Technical Education (AICTE). The institute was established in 2000 and is located at Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore – 560090.

Results – Examples (2/3)

BASELINE

Grigory Neujmin discovered 1147 Stavropolis which has 418476000000.0 apoapsis.

OUR OUTPUT

1147 Stavropolis, a non-family asteroid, was discovered by Georgian-Russian astronomer, Grigory Neujmin at the Simeiz Observatory in the Crimean peninsula. It has an albedo of 0.0466.0 (square meters), a rotation period of 10147.0 and a surface of 13.92 km. It is a rocky surface with an apoapsis of 418476000000.0.

Results – Examples (3/3)

BASELINE

The song Mermaid by Train is written in the reggae music genre and was followed by the band's cover of the John Lennon song Imagine. The original version of Imagine is in the pop music genre and was followed up by Lennon's hit Happy Xmas (War is Over).

OUR OUTPUT

John Lennon is an English musician. He was originally a solo performer of the song 'Imagine' which was published in 1998. During the recording of the song, "Body Counts" he performed it with the guitar and was accompanied by Yoko Ono. The song was published in October 2010 on the B-side of the John Lennon Peace Monument (located in Chavasse Park, Liverpool).

Results – Evaluation (1/3)

Fluency evaluation with SLOR

- SLOR, i.e. syntactic log-odds ratio, is a score for referenceless fluency evaluation of NLG output at the sentence level
- Given a sentence x, SLOR is computed as:

$$SLOR(x) = \frac{\log(p_M(x)) - \log(p_u(x))}{|x|}$$

where

$$p_M(x) = p_M\left(\left(x_1, \ldots, x_{|x|}\right)\right) = p(x_1)\prod_{i=2}^{|x|} p(x_i|x_{<i})$$

and

$$p_u(x) = \prod_{i=1}^{|x|} p(x_i)$$

	WebNLG ref.	KG2Text output	Our output
AVG_{SLOR}	2.79	2.83	3.25

Results – Evaluation (2/3)

Questionnaire-based human evaluation

- Rating of the quality of a small sample of texts according to four dimension:
  1. ***Coherence*** (three options: Yes/No/Somewhat);
  2. ***Grammaticality*** (three options: Yes/No/Somewhat);
  3. ***Faithfulness*** (two options: Yes/No);
  4. ***Informativeness*** (three options: Yes/No/Somewhat)

- For comparison purposes, the baseline is the output of the KG-to-text model
- 13 judges

Results – Evaluation (3/3)

Questionnaires' summary

% (**Baseline**)	Yes	No	Somewhat
Coherence	80.0	6.15	13.85
Grammaticality	82.7	8.08	9.23
Faithfulness	76.15	23.85	-
Informativeness	12.3	78.85	8.85

% (**Our results**)	Yes	No	Somewhat
Coherence	45.38	28.85	25.77
Grammaticality	65.0	18.46	16.54
Faithfulness	35.38	64.62	-
Informativeness	68.08	7.3	24.62

Conclusions

DISCUSSION

- The generated texts have good levels of grammatical correctness and informativeness, but there is room for improvement with regards to textual coherence and faithfulness

FUTURE WORKS

- Large-scale task-specific dataset
- Simpler architectures

Politecnico di Torino

DISMA
ECCELLENZA 2018·2022

Thank you for your attention